Toward Sequencing "Narrative DNA": Tale Types, Motif Strings and Memetic Pathways

Sándor Darányi, Peter Wittek, László Forró†

Swedish School of Library and Information Science

University of Borås

50190 Borås, Allégatan 1, Sweden

†8220 Balatonalmádi, Remetevölgyi út 27, Hungary

Abstract

The Aarne-Thompson-Uther Tale Type Catalog (ATU) is a bibliographic tool which uses metadata from tale content, called motifs, to define tale types as canonical motif sequences. The motifs themselves are listed in another bibliographic tool, the Aarne-Thompson Motif Index (AaTh). Tale types in ATU are defined in an abstracted fashion and can be processed like a corpus. We analyzed 219 types with 1202 motifs from the "Tales of magic" (types 300-749) segment to exemplify that motif sequences show signs of recombination in the storytelling process. Compared to chromosome mutations in genetics, we offer examples for insertion/deletion, duplication and, possibly, transposition, whereas the sample was not sufficient to find inverted motif strings as well. These initial findings encourage efforts to sequence motif strings like DNA in genetics, attempting to find for instance the longest common motif subsequences in tales. Expressing the network of motif connections by graphs suggests that tale plots as consolidated pathways of content help one memorize culturally engraved messages. We anticipate a connection between such networks and Waddington's epigenetic landscape.

Keywords: tale type, motif, motif sequence, mutation, recombination, plot development, memetic pathway, epigenetic landscape

1. Introduction

Recently Darányi (2010) has analyzed the role of formulaity in oral and written narratives, and hinted at a parallel with sublanguages for indexing (Harris, 2002) also used in immunology (Harris *et al*., 1989) and bioinformatics (Leontis & Westhof, 2003). The similarity between these wildly different application domains goes back to the use of motifs. In the literary sense, a motif is an instance of a prominent yet little investigated content-bearing unit: an element that keeps recurring in an artifact – e.g. in film, music, but also in folklore or scientific texts – by means of which often a narrative theme is conveyed. As Uther notes, "Although the definitions of a tale type as a self-sufficient narrative, and of a motif as the smallest unit within such a narrative, have often been criticized for their imprecision, these are nevertheless useful terms to describe the relationships among a large number of narratives with different functional and formal attributes from a variety of ethnic groups, time periods, and genres. The general distinction of a motif as one of the elements of a tale (that is, a statement about an actor, an object, or an incident) is separated here from its content. In fact, a motif can be a combination of all three of these elements, for example, when a woman uses a magic gift to cause a change in the situation. "Motif" thus has a broad definition that enables it to be used as a basis for literary and ethnological research. It is a narrative unit, and as such is subject to a dynamic that determines with which other motifs it can be combined. Thus motifs constitute the basic building blocks of narratives" (Uther, 2004).

On the other hand in bioinformatics oftentimes the task is to compare a protein of unknown structure with its homologues of known 3-D structures based on the idea

of motifs (Buhler & Tompa, 2002). The concept of a motif here refers to a Hidden Markov Model stating that e.g. in a deoxyribonucleic acid (DNA) sequence, amino acids such as arginine, leucine, cysteine and histidine, follow each other with certain probabilities. Based on such conceptual similarities between the two domains, Darányi and Forró (2012) postulate a parallel between coding textual and genetic information, pointing toward "narrative genomics" as a recombination theory of content variation. A related phrase, the concept of "narrative DNA" (i.e. recombinative narrative elements similar to DNA, a building block of life with the genetic instructions used in the development and functioning of all known living organisms) goes back to Bruce (1996), with the idea reinforced by Gill (2011).

This paper is structured as follows: Section 2 discusses related work, whereas Section 3 outlines text evolution as a recombination process. In Section 4 we briefly list the material and method used in this study, with the results in Section 5, their discussion and future work in Section 6, and our conclusions in Section 7.

2. Background considerations and related work

Here we continue to use metadata to exemplify our hypothesis. The metadata in case is the Arne-Thompson-Uther Tale Type Catalog (ATU), a classification and bibliography of international folk tales (Uther, 2004), an alphanumerical, basically decimal classification scheme describing tale types in seven major chapters (animal tales, tales of magic, religious tales, realistic tales (novelle), tales of the stupid ogre (giant, devil), anecdotes and jokes, and formula tales), with an extensive Appendix discussing discontinued types, changes in previous type numbers, new types, geographical and ethnic terms, a register of motifs exemplified in tale types, bibliography and abbreviations, additional references and a subject index.

The numbering of the tale types runs from 1 to 2399 (in fact, 2411). Individual type descriptions uniformly come with a number, a title, an abstract-like plot mostly tagged with motifs, known combinations with other types, technical remarks, and references to the most important literature on the type plus its variants in different cultures. At the same time, as the inclusion of some 250

new types in the Appendix indicates, tale typology is a comprehensive and large-scale field of study, but also unfinished business: not all motifs in the Aarne-Thompson Motif Index (AaTh; Thompson, 1955-58) were used to tag the types, difficulties of the definition of a motif imposed limitations on its usability in ATU, and considerations related to classification of narratives had to be observed as well.[1]

In the ATU, tale types are defined as canonical motif sequences such that motif string A constitutes type X, string B stands for type Y, etc. Also, it is important to note that tale types were not conceived in the void, rather they extract the essential characteristic features of a body of tales from all over the world. An example is an excerpt from Type 300 *The Dragon-Slayer:* "A youth acquires (e.g. by exchange) three wonderful dogs [B421, B312.2]. He comes to a town where people are mourning and learns that once a year a (seven-headed) dragon [B11.2.3.1] demands a virgin as a sacrifice [B11.10, S262]. In the current year, the king's daughter has been chosen to be sacrificed, and the king offers her as a prize to her rescuer [T68.1]. The youth goes to the appointed place. While waiting to fight with the dragon, he falls into a magic sleep [D1975], during which the princess twists a ring (ribbons) into his hair; only one of her falling tears can awaken him [D1978.2]."

Together with the AaTh, ATU is the standard reference tool for librarians and digital curators alike, although other manuals such as Jason (2000) also come handy as means of orientation. When using the ATU, it is regarded as a matter of fact that its descriptive units, motifs, constitute the highest level of abstraction, and there are no units of content above this. However, Darányi and Forró (2012) have recently shown that, contrary to expectations, motifs sometimes agglomerate into higher-order multiplets, some of them being even collocated, i.e. tale types as motif strings are not *entirely* unique and must have been persistent enough to be reused as building blocks of plots.

In the above study, the authors considered ATU as a text corpus and analysed its sub-section "Supernatural adversaries" (types 300-399) in particular and section "Tales of magic" (types 300-749) in general. The two

---

[1] Hans-Jörg Uther, personal communication (02-12-11).

subcorpora were scrutinized for multiple motif co-occurrences and visualized by the two-mode clustering of a bag-of-motifs matrix. Having excluded types not indexed by motifs at all, the first part of the experiment (300-399) worked with 52 tale types defined on the basis of 281 motifs, and the second part (300-745A) with 219

types and 1202 motifs, respectively. After ontology visualization leading to the above conclusion, their cautiously optimistic suggestion was that as the complete AaTh contains about 40.000 motifs, this could allow for the prevalence of robust motif sequences as a new kind of metadata, and enable the use of both single and chained motifs as tags for semantic markup. Secondly, they hypothesized that since only canonical sequences of tale functions (a limited set of action types used by another limited set of actors) are known to result in "valid", i.e. acceptable, Russian fairy tales (Propp, 1968), collocated motif strings might play the same role. Thus motif substrings could be exchanged between narratives in the course of text variation, and a simple model borrowed from genetics, four types of chromosome mutation, could exemplify narrative evolution as a recombination process.

3. Narrative element recombination

These ideas were of interest to us for two reasons. The first broad context was the perception of text variation as an evolutionary process, and the task of mapping evolving semantic content onto structures with both hierarchical and multivariate access. In this frame, the reason why some motif strings have evolved and survived relates to a kind of selection pressure in a cultural historical setting, yet to be modelled. To this end, ATU and AaTh as tools have pioneered and mastered the hierarchical approach to content description but are wanting in terms of being understood as multivariate products at the same time. This is a current deficiency that cannot be overlooked or neglected when it comes to any kind of their overhaul in and for a digital environment.

In other words, for modelling one needs descriptive units of content which can index the source material in its entirety, are both multivariate by nature and fit the hierarchical classification structure, plus flexible enough

to evolve, i.e. become more and more enriched variants of the original standard classifications. Indexing by single text words or phrases plus by motifs is clearly not enough to meet this goal. On the other hand, the existence of persistent motif strings in multiple copies underlying several types indicates that more than one level of semantic metadata may pertain to the body of tales we want to index.

The other broad context is the parallel between the linguistic and the genetic code as vehicles of information transfer over time. Both use coded transfer mechanisms to transmit their messages, capture instructions to reproduce meaning from form (we regard context as form here); and in both, sequence plays an important role in the coding and decoding process.

Tale types as motif sequences follow the sublanguage approach to content representation, pioneered by Harris (2002). As pointed out by Darányi (2010), this domain-specific practice from the life sciences can be recognized in formal descriptions of narrative content, too. Below the similarities between their communication patterns allow for methodology import between the two domains:

(1) Content is sequential, coded by an alphabet and compiled based on the combinations of its elements, i.e. irrespective of their order on a basic observation level. This holds for nucleotides – the building blocks of nucleic acids such as DNA and RNA – and motifs, the building blocks of tale types alike;

(2) On a next level, adding grammar and moving over to permutations, sequences start to play a role. Canonical nucleotide sequences generate secondary and tertiary – in fact spatial – structures such as the famed double helix; canonical motif sequences may contribute to the evolution of tale types, themselves representatives of tale variants. Moreover, function sequences develop into fairy tale subtypes as shown by plot analysis (Propp, 1968), and canonical mytheme sequences constitute myths and mythologies (Lévi-Strauss, 1964-71; Maranda, 2001). In a sense, reading and understanding the genetic code and narratives alike demands the mastering of abstract grammars with their equally abstract vocabularies;

(3) As said the concept of motifs is widely used in bioinformatics. Motifs in this sense mean primary nucleotide sequences of functional importance for

structure generation. Sequential motifs include structural and regulatory motifs, with different functionalities pertaining to them; we anticipate methodological undercurrents linking the two knowledge domains which need to be explored in more detail.

(4) Chromosome and story mutations may be more similar than thought previously. Chromosomal mutations produce changes in whole chromosomes (more than one gene), or in the number of chromosomes present, with the major types being (a) deletion – loss of part of a chromosome; (b) duplication – extra copies of a part of a chromosome; (c) inversion – reversal in the direction of a part of a chromosome; and (d) translocation – part of a chromosome breaks off and attaches to another one.

Whereas most mutations are neutral and have little or no impact on the functionality of the product, their adding up can dramatically affect the survival rate of the outcome, leading to new genotypes and phenotypes in the course of evolution. In the same vein, deletion and translocation could be standard tools in the narrative building toolkit; inversion is suggested to play a central role in the Bible (Christensen, 2003), and duplication is evident e.g. in the case of the Proppian narrative scheme where complete tale moves may be repeated several times or combined with one another by different embeddings (Propp, 1968). This indicates the need for a theory of text evolution as a series of narrative element recombinations, forming from simple to more complex structures by "mutation mechanisms".

4. Material and method

From the sample of 219 tale types as in Darányi and Forró (2012), examples for mutation types were manually selected and disambiguated where more than one tale variant was coded by the same ATU number, plus a set with the same motif (L161) both in terminal and non-terminal positions was separated for network visualization. Until better tools become available and allow for more stringent procedures, we defined insertion and deletion as added or missing inlays within a sequence of motifs. Transposition was considered a single motif or motif string added after a marker. Duplication was regarded as string repetition, and inversion as a reversed motif string.

5. Results

Below we identify three out of the above four major mutation types in our metadata to show how different mechanisms may lead to tale element recombination.

5.1 Insertion and deletion

This type is inherent in e.g. ATU 545A *The Cat Castle*: [B211.1.8 / B422 / B421 / B435.1] - B581.1.2 - **N411.1.1** - F771.4.1 - D711 - **B582.1.**2, and 545B *Puss in Boots*: [B211.1.8 / B422 / B435.1 / B435.2 / B441.1] - [B580 / B581 / B582.1.1] - **K1917.3** - **K1952.1.**1 - [F771.4.1 / **K722**] - D711, where motifs separated by / refer to storytelling alternatives, e.g. both [B211.1.8 / B422 / B421 / B435.1] and [B211.1.8 / B422 / B435.1 / B435.2 / B441.1] represent helpful animals. B581.1.2 and [B580 / B581 / B582.1.1], respectively, stand for bringing luck; F771.4.1 is castle owned by ogre, and D711 means disenchantment by decapitation. Therefore the underlying joint storyline is "Helpful animal brings luck by defeating ogre, culminating in his own decapitation". In the first plot, N411.1.1 (Cat as sole inheritance) and B582.1.2 (Animal wins husband for mistress) are insertions indicated by boldtype, whereas in the second, K1917.3 (Penniless wooer: helpful animal reports master wealthy and thus wins girl for him), K1952.1.1 (Poor boy said by helpful animal to be dispossessed prince (wealthy man) who has lost clothes while swimming (in shipwreck)), and K722 (Giant tricked into becoming mouse. Cat eats him up) appear as additions to the basic plot. Since in *The Cat Castle*, a poor girl finds a husband, whereas in *Puss in Boots*, a poor man marries a princess, i.e. we have the heroine and hero oriented variants of the same story, it is an open question whether additions or deletions have resulted in these variants.

5.2 Transposition

For transposition, we depart from the observation that in the sample, motif L161 (Lowly hero marries princess) occurred in 20 tale types (9 % of the 219 plots), and out of these, it was in 15 types in terminal position, i.e. the tale finished with the wedding, whereas in 5 cases the adventures continued.

Consider the story of Aladdin as an example. Its ATU summary goes like this: "A magician orders a (stupid)

boy, Aladdin, to fetch a lamp for him out of a cave of treasures. The cave opens and closes by means of a magic ring [D1470.1.5]. Aladdin finds the lamp [D812.5, D840, D1470.1.16, D1421.1.5, D1662.2], but when he wants to leave the cave it does not open (the magician has closed it). When Aladdin rubs the magic ring (lamp) in despair, a helpful genie appears and leads him out. Aladdin reaches his mother's house and wishes for riches and a castle [D1131.1]. Both wishes are fulfilled by the genie (by another spirit who appears in the same way when the lamp or the ring is rubbed). Aladdin woos the princess, but her father intends to marry her to another man (Aladdin marries the princess [L161]). The magician exchanges the old, magic lamp (which the princess had kept) for a new, worthless one [D860, D371.1]. He wishes himself to be transferred to Africa together with the princess and the castle [D2136.2]. Aladdin is imprisoned. He rubs the ring [D881] and the genie takes him to the castle where the princess is. She poisons the magician (Aladdin kills him). Aladdin takes the lamp again and uses it to return with the castle and the princess to his home." One can easily anticipate a tale variant which finishes with the wedding, so that a second, from somewhere else translocated plot could be concatenated. This is described as: 561 *Aladdin*: D1470.1.5 - [D812.5 / D840 / D1470.1.16 / D1421.1.5 / D1662.2] - D1131.1 - L161 - [**D860 / D371.1**] - **D2136.2** - **D881**.

The other three examples are as follows:

Type 502_1 *The Wild Man*: "A king catches a wild man (Iron John) and puts him into a cage, forbidding anyone to set him free. His son frees the prisoner because his ball rolls into the cage or because he feels pity for him. The prince is afraid of his father's anger and leaves home (his father drives him away to be killed or sends him to another king) along with a servant. On their way the servant persuades the prince to exchange clothes. The prince becomes a servant at the court of another king. At a tournament he appears unrecognized three times on a splendid horse [R222] which he received from the wild man and wins the hand of a princess. Or, he wins the princess because he has helped her father in war [L161]. Often the wild man is disenchanted [G671]. In some variants the prince works for a while at the wild man's house where he disobeys instructions (e.g. looks into a forbidden chamber [C611], cares for a horse although it

is not allowed [B316]) and his hair turns to gold." As this is a tale whose initial situation is not formalized in terms of motifs, we summed up the plot of the first variant as R222 - L161 - **G671**. In its second variant, no wedding takes place, i.e. L161 is missing, hence that version was not considered for exemplification here. However, in the above variant, G671 as a new ending to the story suggests a possible transposition.

Type 400_1 *The Man on a Quest for His Lost Wife* is summed up in the ATU as follows: "This tale exists chiefly in three different forms: (1) A man in distress (impoverished fisherman, merchant) unwittingly promises his (unborn) son to the devil [S240]. When the boy is delivered to him later, the devil cannot use him because he is protected by magic [K218.2] (cf. Type 810). Thus the boy is cast out in the sea (river, desert). He arrives in a foreign country and finds a lonely castle where he meets a bewitched princess (maiden, fairy) in the form of a serpent (deer). He rescues her by enduring three nights of torture [D758.1]. They marry [F302, L161]. When he wants to visit his parents, his wife gives him a ring to carry him home [D1470.1.15], and she forbids him to call her to come to him [C31.6] (to boast of her beauty [C31.5]). At home he is induced (by his mother) to break the taboo. His wife appears [D2074.2.3.1], takes the ring, and leaves him destitute. The man sets out in search of his wife [H1385.3]. On his way he meets three hermits (rulers of animal kingdoms, or moon, sun, and wind) whom he asks for directions [B221, H1232, H1235]. With the help of the third he arrives at the empire of his wife, or he pretends that he wants to help three giants who are fighting over magic objects (inheritance, booty). He steals the magic objects (magic sword [D1400.1.4], magic coat or hood [D1361.14], seven-league boots [D1521.1]) [D831, D832] (cf. Type 518). With their help he is able to overcome the obstacles on the way to his wife [D2121]. When he finds his wife, she is about to marry another man [N681]. He discloses his identity as her real husband.(2) Meeting the princess and disenchantment as in version (1); but the disenchantment is not complete. The princess wants to travel back to her own distant land. She asks her rescuer to wait for her at a certain time and place. She appears three times, but each time a servant (witch) has put her husband into a deep sleep

from which he cannot be awakened [D1364.15, D1364.4.1, D1972]. The princess informs him (in a letter) how and where to find her (on the glass mountain). The man sets out to find her. Continued as in version (1). (3) A youth watches a flock of birds (swans, ducks, geese, doves) land on the shore. The birds take off their feather coats and become beautiful maidens [D361.1]. While they are bathing, the youth steals the feather coat of the most beautiful girl, who cannot leave with the others and thus must marry the youth [D721.2, B652.1]. Later, because of carelessness (of the man's mother), the maiden takes back her coat [D361.1.1] and flies away (together with her children). She tells the youth her destination in the otherworld (e.g. glass mountain). The man sets out in search of his wife (as in version 1)." As for the three variants, the formula of the first one is: S240 - K218.2 - D758.1 - [F302 / L161] - **D1470.1.15 - [C31.6 / C31.5] - D2074.2.3.1** - H1385.3 - B221 - H1232 - H1235 - D1400.1.4 / D1361.14 / D1521.1 - [D831 / D832] - D2121 - N681. The second one, 400_2, replaces the segment **D1470.1.15 - [C31.6 / C31.5] - D2074.2.3.1** by **D1364.15 - D1364.4.1 - D1972** which is regarded transposition for the time being, and repeats the rest of the string from H1385.3 to N681. The third variant, 400_3, mentions that the beautiful girl having lost her bird shape must marry her captor but does not index the story with L161, nonetheless after having replaced the beginning of the plot by **D361.1 - [D721.2 / B652.1] - D361.1.1**, i.e. bird shape lost and regained, it continues with H1385.3 to N681 as above.

Finally type 303 *The Twins or Blood-Brothers* tells the following story: "After having eaten a magic fish (apple, water) [T511.5.1, T511.1.1, T512], a woman gives birth to twins. (Cf. Type 705A.) Grateful animals accompany the grown-up brothers, or animals give them one or more of their young ones because the brothers did not kill them. (The brothers are given unusual animals; they win them or bring them up; in some variants, the animals are born at the same time as the brothers [T589.7.1].) Together with his animals, one of the brothers sets out. When the brothers separate, they agree upon a life token that gives a warning when one of them is in mortal danger and needs help: Water will become cloudy, a plant or a tree dry up, a knife stuck in a tree will grow rusty, etc. [E761]. The first brother frees a princess (three princesses) from a dragon (trolls), unmasks an impostor ("Red Knight") who pretended to be the princess's rescuer, and marries the princess [R111.1.3, K1932, H83, L161]. Cf. Type 300. Against a warning, the hero follows a light [G451] (is tempted by an animal). He falls into the power of a witch and is turned to stone [D231]. His twin brother is warned by the life token and sets forth in quest of him. The princess mistakes him for her husband, as the two brothers are very much alike [K1311.1]. At night the brother puts a naked sword in the bed between himself and his sister-in-law [T351]. Then he finds the witch, makes her remove the spell from his brother, and kills her. The first brother learns that the second has slept with his wife and kills him out of jealousy [N342.3]. Later on, when he asks his wife why she had put the sword in the bed, he realizes that his brother was innocent. The brother is resuscitated by magic means [B512] (water of life). In some variants, a youth saves the life of a raven (crane, eagle). As a reward he obtains magic objects. The youth defeats a sea monster, delivers three princesses, and marries the youngest of them." Its formula is: [T511.5.1 / T511.1.1 / T512] - T589.7.1 - E761 - R111.1.3 - K1932 - H83 - L161 - **G451- D231- K1311.1 - T351 - N342.3- B512**. We regard the segment in boldtype as a transposition but at the same time warn the reader that screening for the transposed chunks in the complete ATU was not possible for this paper, and therefore this part of our results remains a suggestion only (Table 1).

Finally, for the same reason, we were not able to isolate inversion, i.e. reversed motif string in our material.

[Table 1 comes approximately here]

5.3 Duplication

A good example for motif string duplication is type 700 *Thumbling*: "A childless couple wish for a child, however small he may be. They have a boy (by supernatural birth) the size of a thumb [F535.1]. Thumbling takes food to his father on the farm and drives the wagon (plow) by sitting in the horse's (ox's) ear [F535.1.1.1]. He allows himself to be sold to strangers and then runs away from them. He lets himself be sold to thieves and accompanies them while they steal. Thumbling either helps them or he betrays them by

his shouting; he then robs the thieves. Cf. Type 1525E. He is swallowed by a cow [F911.3.1], talks from the cow's insides and reappears [F913] (in the sausage prepared from the intestines of the slaughtered cow [F535.1.1.8]). Someone takes the intestines (sausage) and, frightened by Thumbling's voice inside, throws them away. Thumbling is swallowed by a wolf (fox) who eats the intestines [F911.3.1]. He talks from the wolf's belly and the wolf becomes sick and frightens (warns) shepherds. The wolf dies (is killed) and Thumbling is rescued [F913], or he persuades the wolf to take him to his father's house [F535.1.1]." We notice that in the respective motif sequence, F535.1 - F535.1.1.1 - **F911.3.1** - **F913** - F535.1.1.8 - **F911.3.1** - **F913** - F535.1.1, the segment in boldtype is repeated twice.

It is interesting to compare *Thumbling* with the related type 333 *Little Red Riding Hood*: "A little girl, called "Red Riding Hood" because of her red cap, is sent to her grandmother who lives in the forest and is warned not to leave the path [J21.5]. On the way she meets a wolf. The wolf learns where the girl is going, hurries on ahead, and devours the grandmother (puts her blood in a glass and her flesh in a pot). He puts on her clothes and lies down in her bed. Red Riding Hood arrives at the grandmother's house. (She has to drink the blood, eat the flesh, and lie down in the bed.) Red Riding Hood doubts whether the wolf is her grandmother and asks him about his odd big ears [Z18.1], eyes, hands, and mouth. Finally the wolf eats Red Riding Hood [K2011]. A hunter kills the wolf and cuts open his belly. Red Riding Hood and the grandmother are rescued alive [F913]. They fill the wolf's belly with stones [Q426]; he is drowned or falls to his death." Its formula is J21.5 - Z18.1 - K2011 - F913 - Q426, that is, both tales contain the motif F913./Victims rescued from swallower's belly/ (Table 2). Representing now the two related tale types as a directed graph whose nodes stand for the motifs and whose edges are numbered according to tale types, we notice that motif duplication yields a loop (Fig. 1).

[Table 2 and Figures 1-2 come approximately here]

5.4 Plots as memetic pathways

For visual inspection we regarded the motif index of ATU as a description of a directed graph whose nodes are motifs from AaTH. A directed edge starts from motif A to motif B if there is at least one tale type in which motif A and motif B are subsequent motifs in this order. An edge is labelled by all the tale types in which such an order appears (Fig. 2).

We note in passing that tales and their variations have been created by thousands of individuals, which is also true for content on the World Wide Web. While individuals can impose order on the web at the local level, its true global organization is utterly unplanned, and high-level structure needs to be extracted a posteriori. If we consider the graph of the web where the nodes are websites and the directed edges are links between them, we may notice the presence of so-called hubs and authorities (Kleinberg, 1999). A hub is a page that points at many other pages, whereas an authority is a page that is linked in by many different hubs.

Google's PageRank algorithm followed this line of thought to evaluate websites and rank websites (Page *et al*., 1999). Trying to establish a ranking of motifs, we attempted to find a similar structure in their network. Early results however remained inconclusive and indicate the absence of clear hubs and authorities in our limited sample. There are motifs with a high number of both incoming and outgoing edges, but no definite sinks or sources. Therefore a ranking will have to be based on centrality or the degree of a node.

This is illustrated in Fig. 2 where, even at this small scale, the motif network shows an interesting structure. For example K1932 makes an excellent dense centre which exemplifies that there are no real hubs or authorities, but common motifs that appear in different tale types and in different positions. A hub would mean a motif from which there is an extraordinary number of possible continuation in different tale types. We do not see this, therefore we may believe that story lines follow a restricted number of possibilities (hence one can rightfully suspect a Hidden Markov Model). An authority or a sink in the graph would be a motif that gathers plot lines, many different tale types would end or go through the very same motif. We do not see this either. H1242 is similar to K1932.

Another interesting option is to depart from the

engraving function of storylines. When repeated in the course of oral transmission, such as retelling, such canonical plots as tale types preserve themselves by being repeated a thousand times and resulting in as many variants. With the above graph representation convention, one is in a position to combine this engraving function and the now forking then intertwined nature of the web of plots with individual storylines as memetic pathways. Memetic refers here to memes, those hypothetical units of cultural heritage which, by analogy with genes, self-replicate to maintain themselves (Dawkins, 1976). In this somewhat lose analogy, self-replication errors in genes result in mutations whereas self-replication errors in memes lead to text variation.

6.   Discussion and future work

Our ongoing experiments suggest that better algorithms will identify not only motif sequences, but will also yield visual representations of the major "narrative mutation" types. In other words we expect that by visual inspection of a network of memetic pathways, one will be able to tell apart more popular motifs from less used ones, plus spot characteristic narrative element recombinations underlying ATU.

Secondly, by considering plot direction as its gradient, we anticipate a connection between such pathways and Waddington's epigenetic landscape (1957). Brock explains the significance of this concept as follows: "Genes provide continuity and a degree of permanence, passing in predictable ways from parents to offspring, from cell to dividing cell. Genes can be detected and sequenced, their frequencies quantified. Much more elusive, though, are the effects of environment on genes. Remarkably, in 1932, at a time when genes were recognized as discrete heritable units but their structure and function unknown, Conrad Hal Waddington used the term 'epigenetics' to refer to the external manifestation of genetic activity. He presented the 'epigenetic landscape' as a way to visualize the forces affecting cell differentiation. In this model, marbles (cells) move varying ways down a landscape whose contour is affected by genes. Details within the contours are further defined by factors above ('epi-') the fixed genetic level, and these details determine the final resting state of differentiation for each cell type.

Whether epigenetic factors act above, below, before, or after the gene depends on the factor. More importantly, 'epigenetics' today commonly refers to changes that are heritable but do not involve changes in the DNA sequence. Specifically, these are changes that affect gene expression, without changing DNA sequence, which can be passed on at least one generation" (Brock, 2010). As far as we can tell, Waddington's original idea could model the interaction between motifs shaping a landscape from "below" in a tectonic sense, socio-historical constraints influencing it from "above", and plot development as the marbles rolling down the landscape, while its modern interpretation would possibly amount to different readings of the same storyline without alterations to its narrative structure.

The formal connection between memetic pathways and the epigenetic landscape is that two-dimensional (planar) graphs correspond to landscapes (Cantwell & Forman, 1993; Minor & Urban 2008).

7.   Conclusions

To use the terminology of Dawkins (1976), we considered tale types as memetic sequences of motifs, i.e. semantic content with a memory engraving function. Carried out manually, an initial tale type screening on a small test sample indicated that insertions, deletions, repetitions and possible transpositions of single motifs or motif sequences in the sample metadata corpus were not unlike chromosome mutations in genetics.

To regard the development of sequential semantic content an evolutionary process will have to be addressed in more detail in a next paper. Just identifying common structure between tales, and variation in such structure is not sufficient to claim evidence for evolution though. The problem of handling text variation has been there since the 19th century, and regarding text variants as an evolutionary series goes back to Lévi-Strauss' Oidipus analysis (1958) and his consecutive research on the canonical formula of myth. Hence the genetic metaphor for storytelling is a clarification attempt to see if one can model the process to a better extent, and the term "evolution" was used in a loose sense, indicating some sort of directed progress, just like e.g. in cultural evolution. It is also clear that a fitness function will be crucial to prove our point but we focused on simpler

parts of the proposed model at this time.

It remains to be seen if motif networks based on more material than our current sample will show the hubs and authorities structure of the web. Our current assumptions are based on the analysis of a much larger graph whose visualization for this paper ran into problems hence we regard this issue unresolved. However, in another paper we report about adding taxonomy-like information to see if a more explicit graph structure will result (Declerck *et al*., 2012). Finally, to map natural language expressions in tales to motifs as higher order content indicators, i.e. actively incorporate features at the fine-grained, grammatical level of folk narratives remains a critical task (Lendvai *et al*., 2010).

8. Acknowledgements

9. References

Brock, T. (2010). Nurturing the Concept of Epigenetics. Cayman Chemicals Technical Report 2150 (April 2010).

Bruce, W.K. (1996). Splicing the double helix: Narrative DNA and system assaults. Muncia: Ball State University.

Buhler, J., Tompa, M. (2002). Finding Motifs Using Random Projections. *Journal of Computational Biology* 9 (2), pp. 225-242.

Cantwell, M.D., Forman, R.T.T. (1993). Landscape graphs: Ecological modeling with graph theory to detect configurations common to diverse landscapes. *Landscape Ecology* 8(4), pp. 239--255.

Christensen, D. L. (2003). *The unity of the Bible: exploring the beauty and structure of the Bible.* Mahwah, N.J.: Paulist Press.

Darányi, S. (2010). Examples of Formulaity in Narratives and Scientific Communication. In *Proceedings of the 1st International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts*. Szeged: University of Szeged, pp. 29-35.

Darányi, S. and Forró, L. (2012). Detecting Multiple Motif Co-occurrences in the Aarne-Thompson-Uther Tale Type Catalog: A Preliminary Survey. *Anales de Documentación*, In press.

Dawkins, R. (1976). *The Selfish Gene*. Oxford: Oxford University Press.

Declerck, T., Lendvai, P., Darányi, S. (2012). Multilingual and Semantic Extension of Folk Tale Catalogues. Accepted for DH-12.

Gill, J. (2011). An investigation of cultural complexity via memetics: Methodological rationale and its operationalisation. In: Sheffield Doctoral Conference, Sheffield, 19-20th April 2011. (Unpublished).

Harris, Z.S., Gottfried, M., Ryckman, T., Mattick, P., Daladier, A., Harris, T.N., Harris, S. (1989). *The form of information in science: analysis of an immunology sublanguage.* Dordrecht: Kluwer.

Harris, Z. S. (202). The structure of science information. *Journal of Biomedical Informatics* 35, pp. 215–221.

Jason, H. (2000). *Motif, Type and Genre. A Manual for Compilation of Indices & A Bibliography of Indices and Indexing,* Helsinki: Academia Scientiarum Fennica.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46, pp. 604-632.

Lendvai, P., Declerck, T., Darányi, S., Gervás, P., Hervás, R., Malec, S., Peinado, F. (2010). Integration of Linguistic Markup into Semantic Models of Folk Narratives: The Fairy Tale Use Case. In: *Proceedings of the Seventh International conference on Language Resources and Evaluation*, Pages 1996-2001, Valetta, Malta, European Language Resources Association (ELRA), pp. 1996-2001.

Leontis, N.B., Westhof, E. (2003). Analysis of RNA motifs. *Current Opinion in Structural Biology*, 13, pp. 300-308.

Lévi-Strauss, C. (1958). The structural study of myth. In Sebeok, T.A. (Ed.): *Myth: A Symposium.*. Bloomington: Indiana University Press, pp. 50-66.

Lévi-Strauss, C.(1964-71). *Mythologiques I-IV.* Paris : Plon.

Maranda, P. (Ed.). (2001). *The double twist: from ethnography to morphodynamics.* Toronto: University of Toronto Press.

Minor, E.S., Urban, D.L. (2008). A Graph-Theory

Framework for Evaluating Landscape Connectivity and Conservation Planning. *Conservation Biology*, 22 (2), pp. 297–307.

Page, L., Brin, S., Motwani, R., Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Stanford InfoLab, 1999.

Propp, V.J. (1968). *Morphology of the folktale*. Austin: University of Texas Press.

Rodolfa, K.T. (2008). Inducing pluripotency. StemBook (Ed.), *The Stem Cell Research Community*, StemBook.

Thompson, S. (1955-58). *Motif-Index of Folk-Literature 1–6*. Bloomington: Indiana University Press.

Uther, H.J. (2004). *The Types of International Folktales: A Classification and Bibliography. Based on the System of Antti Aarne and Stith Thompson*, Part I, Helsinki: Academia Scientiarum Fennica.

Waddington, C.H. (1957). *The Strategy of the Genes; a Discussion of Some Aspects of Theoretical Biology* London: Allen & Unwin.